CW01 - Analysing UK Health Profiles

Task 1- Overview - Analysis Strategy

The purpose of this analysis strategy is to clearly outline a logical process such that the initial requirements from UK authorities can be accomplished. The requirement is to inform the health authorities on the relations between health indicators and socio-economic and demographic indicators. Through the use of the exploratory data analysis techniques the intent is to understand and gain insight into the data and to draw out conclusions and develop models which identify the relationships between variables. The steps at a high level have been summarised below:

- Data Collection
- Preparation of the Data for Analysis
- Exploratory Data Analysis
- Development of Models and Visualisation

Data Collection
The collection of the data has already been completed by the local authorities and is stored remotely on a web server. The data is in the form of thirty two individual files (both CSV and XLSX versions exist) and the first phase will be to merge all of the datasets together; this will provide an easier means of comparison as the data will be in one concise data frame.

Preparation of Data for Analysis
Once the data has been merged into a master data frame that includes the name of the Area, ONS codes and each of the thirty two indicators, the preparation can begin. Inspecting that the data has been merged accurately is essential before continuing, this can also be confirmed by getting the information of the data frame as an additional confirmation. It is essential to identify missing values within the data as this could have an-effect on any transformation that needs to be done. Additionally, missing values could lead to inaccurate predictions; therefore a decision of how to deal with these must be made. Finally, any transformations of the data need to be applied, this may be to normalise the data or alternately the use of dimension reduction (PCA) to consolidated the variables. This can be achieved through thorough inspection of the descriptive statistics, complemented by precise visuals allowing the ability to make accurate decisions on how to deal with the data.

Exploratory Data Analysis
At this stage the data is in an appropriate state to analyse. This analysis will be completed using data mining techniques (and basic tests of the data) to identify outliers and to understand the correlations between variables and basic tests of the data. These tests could include testing for normality, correlation (Spearman, Pearson) and t-tests for significance of the chosen variables. Iteratively moving through the previous stage and this stage depending on the output of the analysis when exploring the data.

Development of Models and Visualisation
Once the background work has been completed, the development of statistical models can be commenced. The types of models to be developed are dependent on the analysis completed up to this point, however the model/s used must work towards meeting the objective set at the start helping the authorities understand the relationships between the variables in their data. Once a model has been generated, the next stage is to present representative visualisations which accurately convey the messages which have been identified by the model.

The table below describes what type of variable each of the categories was identified as:

Indicator Category	Dependent / Independent
Our Communities	Dependent
Children's and young people's health	Independent
Adults health and lifestyle	Dependent
Disease and poor health	Independent
Life expectancy and causes of death	Dependent

The dependent categories are dependent on a variety of measures, namely societal status and services available in the authorities,whereas the independent categories have no immediate dependencies.

Dependent variable:

1. Deprivation

Independent variables:

1. Drug Misuse
2. Acute sexually transmitted infections
3. Obese Children (Year 6)

Task 2 – Prepare Data

In the script I chosen to load all of the CSV files in a directory into one master data frame. The chosen method will loop through each of the files selecting only the relevant columns: for the first files Area Name, ONS Code and Indicator Values will be taken; for subsequent files only the Code and Indicator Value will be taken. The reason for this is the ONS code is a unique key between all of the datasets. I have chosen to merge using an outer join to eliminate any potential missing ONS values in the sheets [*Lines 29-92*]. The next steps ensure that the data has been loaded correctly and will evaluate some simple statistics and information on the data frame [*Lines 96-116*].

The next phase was to identify any possible missing values if they existed, from inspection it was immediately possible to identify three columns that contained missing values; Acute Sexually Transmitted Diseases, Incidence of Malignant Melanoma and Starting breast feeding [*Lines 119-197*]. For the first indicator - Acute Sexually Transmitted Diseases – I found that there were two areas with missing values that were in close geographic proximity to each other; Devon CC and West Devon WC. For these missing values I decided to find the mean of the local area and apply those to missing values. It seemed important to note that the National Average varied from the Local Average for this indicator.

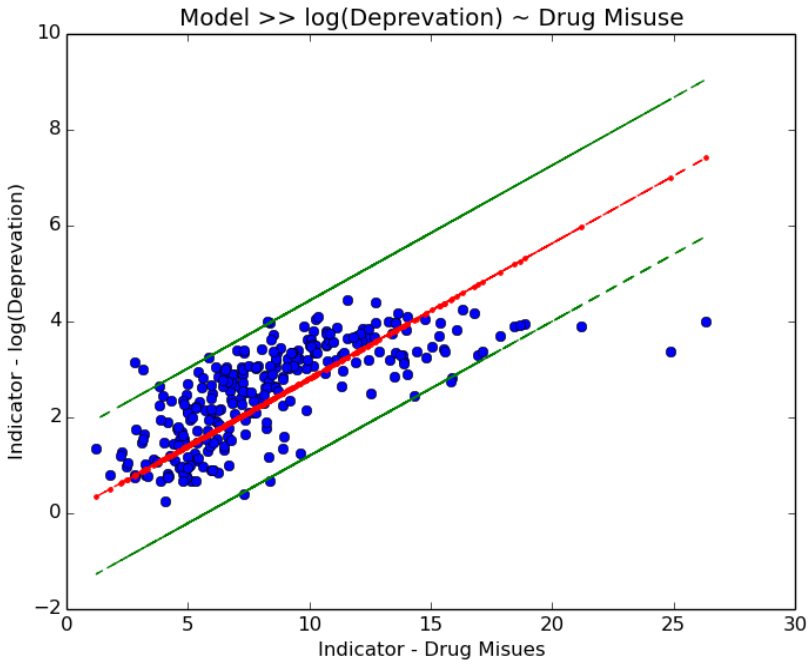
The second indicator, Incidence of Malignant Melanoma, included one missing value in the area called ‘Tower Hamlets LB’. Rather than ignore this missing value I considered the figure for the London area and surrounding areas. Two properties were noticeable; firstly, the London average was lower than the rest of the country and secondly, the surrounding areas were considerably lower than the London average, therefore the replacement value used for Tower Hamlets LB was the average of all the local areas. Finally for the last indicator, Started Breast Feeding, the missing values were not localised therefore the same logic that had been previously used could not be applied. From inspection of the histogram [*Appendix 1.14*], the data was found to be slightly positively skewed with some potential outliers in the lower regions. Despite this, the mean was used to identify the missing value as the distribution of the data was approximately normally distributed.

In the next phase I visualised the data by each category. For each category I produced a scatter matrix to see the distributions of each of the indicators within the categories and also the correlation between the variables [*Appendix 1.3-1.7*]. Additionally, I made comments on each of the indicators where the most repeatedly the variables the indicators were skewed. For each of the categories a Kolmogorov–Smirnov test was used to test goodness of fit to the normal distribution, there was no indication that any category had a P-Value greater than 0.05 when tested, which leads to the conclusion that they are all normally distributed[*Lines 206-272*].

The only transformation applied to the data was a log transformation; the purpose of this was to reduce the effects of skew from the Indicators. Since some of the columns contained zero values, this needed to be considered. A very small constant of 1 was added to the row before taking the logarithm of the value, this avoided infinities arising and minimised negative values. The decision to add 1 as opposed to a smaller constant was because after visualising and performing analysis on the data in Part 3 it seemed a negative value made the data for deprivation misleading [*log(1) = 0*].

Task 3 – Perform Analysis

From the analysis of the dependant and independent variables it is possible to conclude that with statistical significance it is possible to predict the level of deprivation in an area using the ratio of drug use. This can be seen from the plot shown below:



The model that has been created concludes that ~~is~~ the number of cocaine users is a good predictor of the level of deprivation in a given area. The accuracy of the model is high with regards to the variation explained by it [91.1%], this indicates a very good fitting model. From the Analysis of Variance [*Appendix 1.14*], other key indicators also support the model selection as they are also very high – AIC, the model P-Value, and the P-Value of the variable. Therefore with a good level of confidence it is possible to use the model to forecast what the current level of deprivation in a local area might be.

When conducting the analysis of the data there was variation in the data that would be relevant to the authorities as they different significantly from the rest of the Areas in the dataset [*Appendix 1.15*]. Where a label of “one” in the adjacent column identifies the values that are outliers.

Appendix:

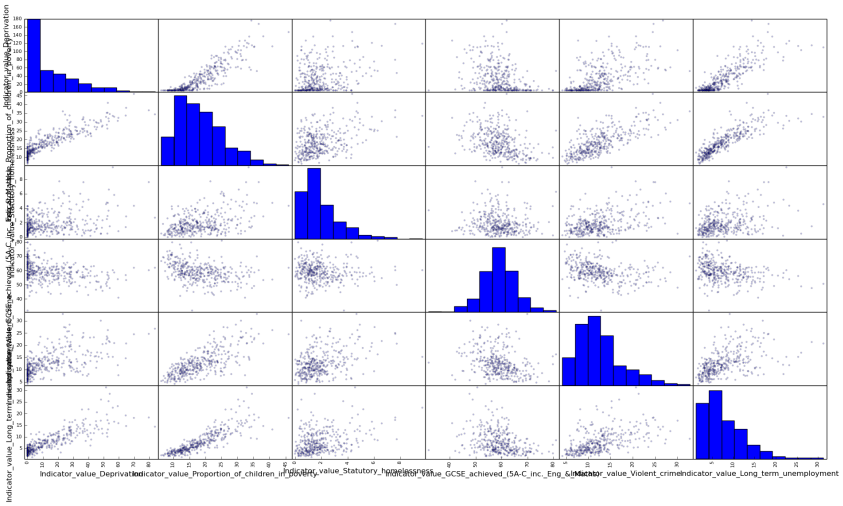
Appendix 1.1 - Correlation coefficient between variables:

Pearson Correlation Coefficient between Indicator_value_Deprivation and Indicator_value_Drug_misuse:	0.7624
Pearson Correlation Coefficient between Indicator_value_Deprivation and Indicator_value_Acute_sexually_transmitted_infections:	0.5722
Pearson Correlation Coefficient between Indicator_value_Deprivation and Indicator_value_Obese_Children_(Year_6):	0.7006

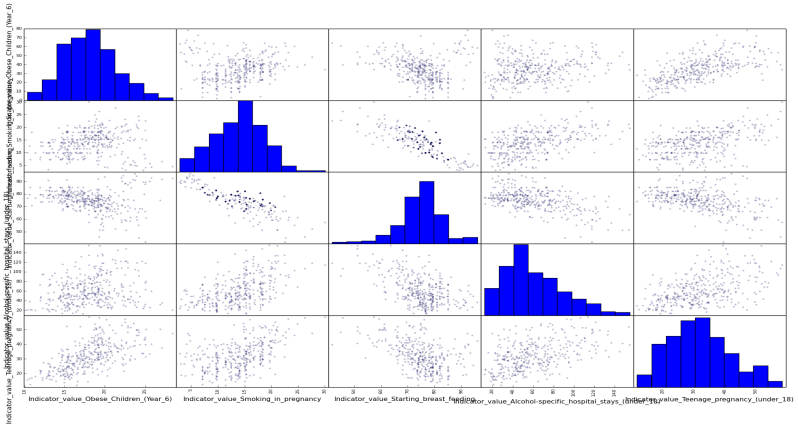
Appendix 1.2 - Number of Outliers Values:

isOutlier - Indicator_value_Deprivation	0
isOutlier – Indicator_value_Drug_misuse	15
isOutlier – Indicator_value_Acute_sexually_transmitted_infections	22
isOutlier – Indicator_value_Obese_Children_(Year_6)	18

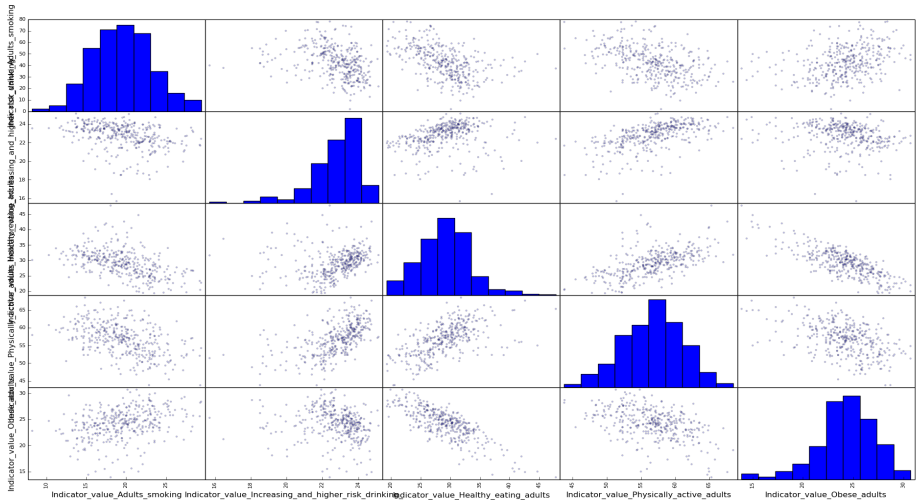
Appendix 1.3 - Our Communities:



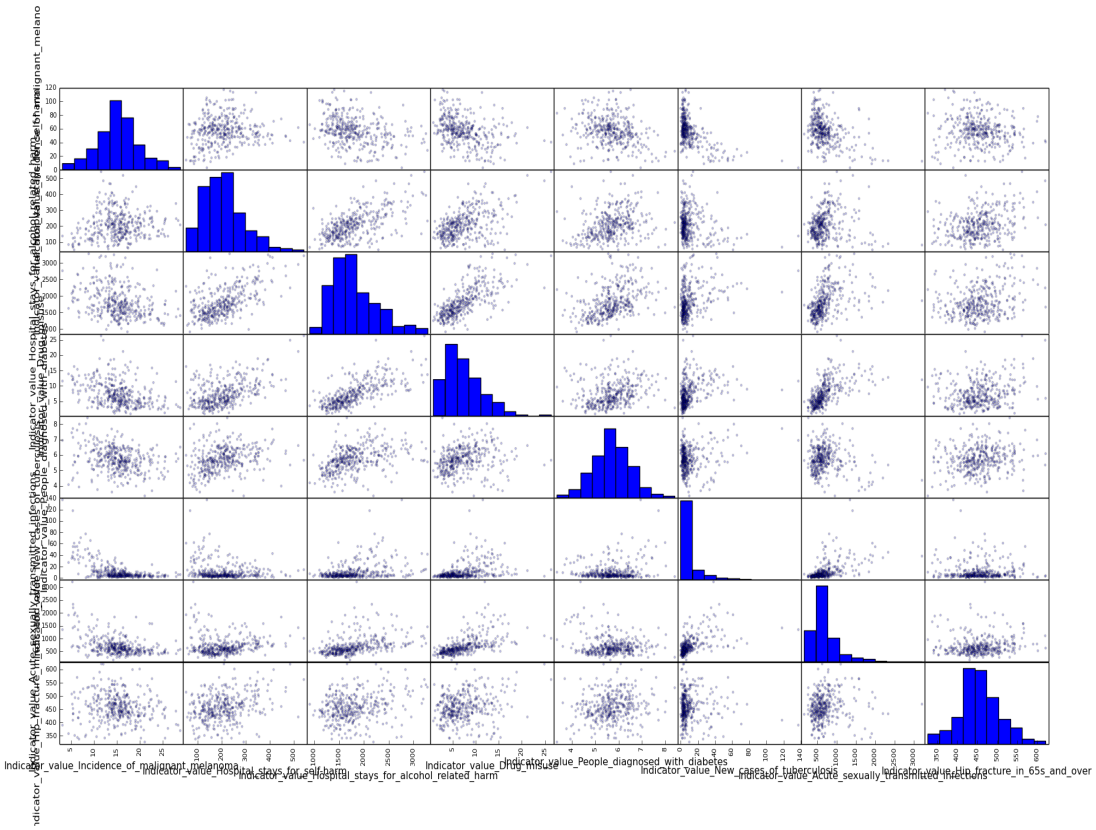
Appendix 1.4 – Children's and Young People's Health:



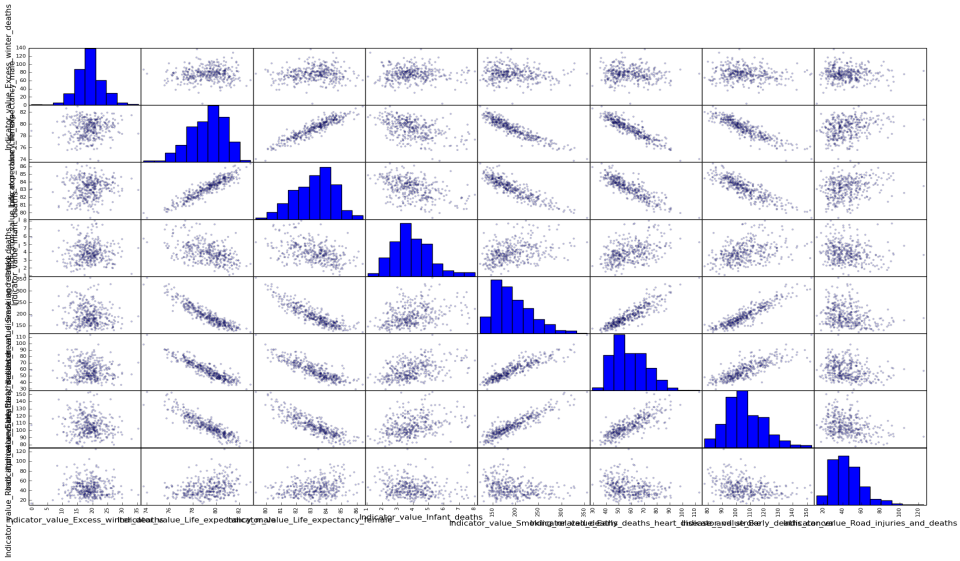
Appendix 1.5 - Adults' Health and Lifestyle:



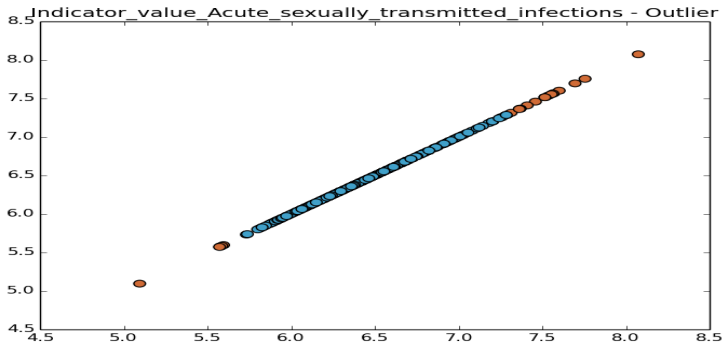
Appendix 1.6 – Disease and Poor Health:



Appendix 1.7 – Life Expectancy and Causes of Death:

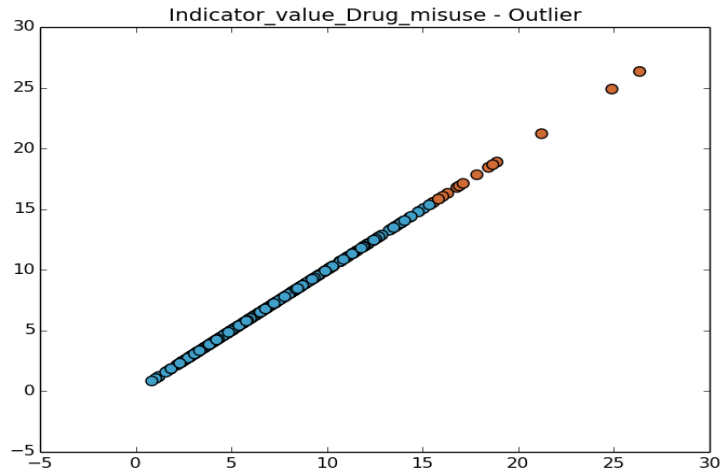


Appendix 1.8 – Indicator Acute Sexually Transmitted Infections:



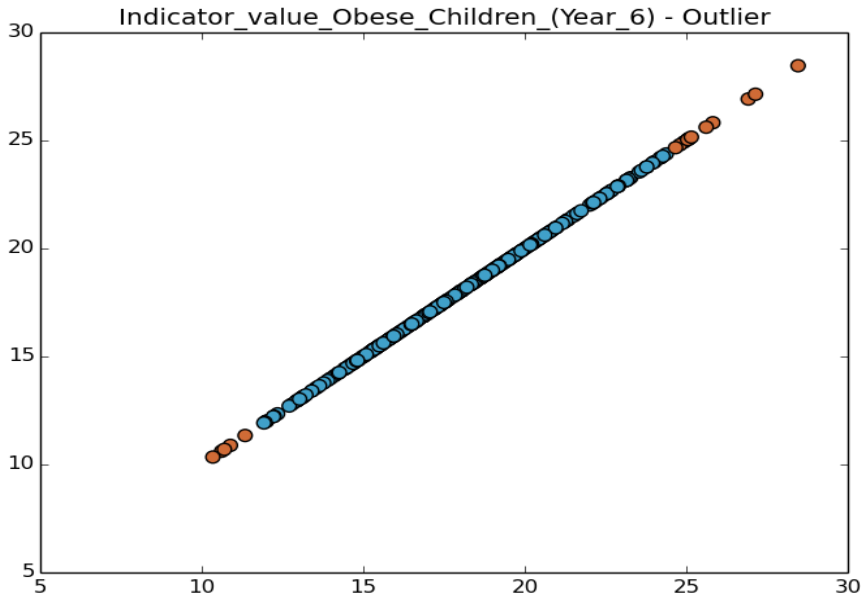
Outlier =Orange

Appendix 1.9 – Drug Misuse:



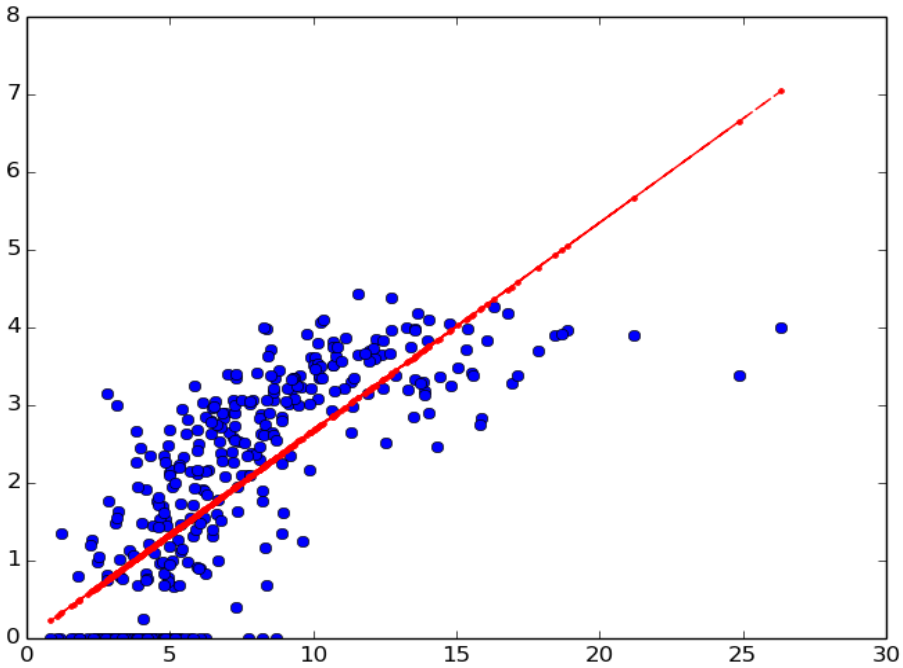
Outlier =Orange

Appendix 1.10 – Indicator Obese Children:

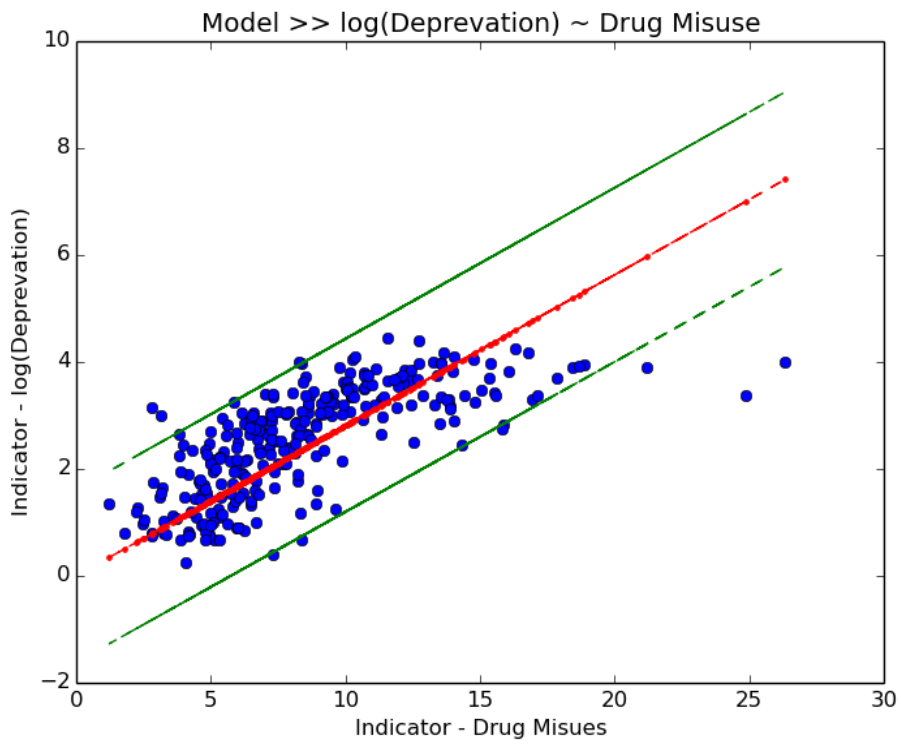


Outliers – Orange

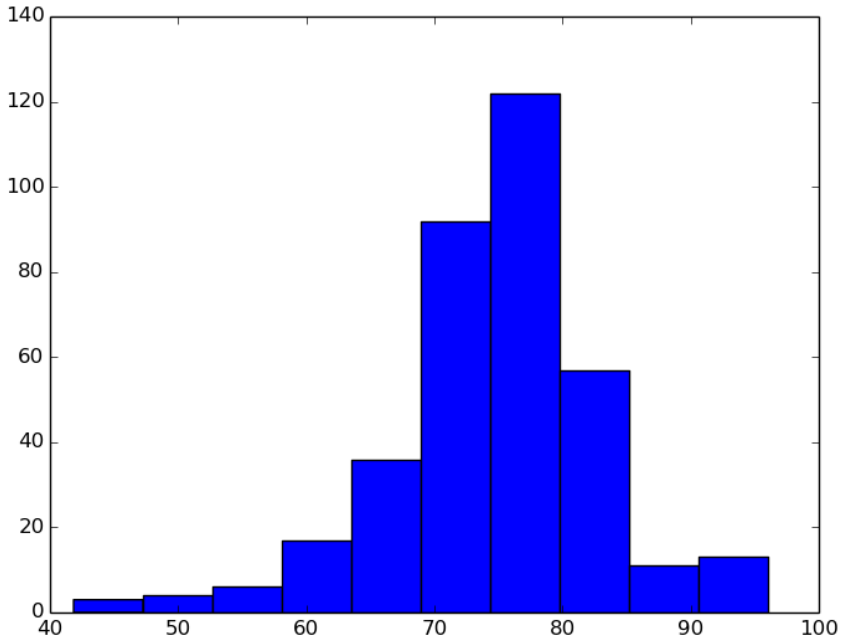
Appendix 1.11 – Model (Model >> Indicator_value_Deprivation = 0.267455Indicator_value_Drug_misuse):



Appendix 1.12 – Model (Model >> Indicator_value_Deprivation = 0.281523Indicator_value_Drug_misuse):



Appendix 1.13 – Distribution of Indicator – Started Breast feeding:



Appendix 1.14 – Analysis of Variance Summary Table

OLS Regression Results

Dep. Variable:	y	R-squared:	0.911
Model:	OLS	Adj. R-squared:	0.911
Method:	Least Squares	F-statistic:	2893.
Date:	Sat, 15 Nov 2014	Prob (F-statistic):	2.75e-150
Time:	21:26:49	Log-Likelihood:	-344.88
No. Observations:	283	AIC:	691.8
Df Residuals:	282	BIC:	695.4
Df Model:	1		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
x1	0.2815	0.005	53.785	0.000	0.271	0.292

Omnibus:	55.653	Durbin-Watson:	1.742
Prob(Omnibus):	0.000	Jarque-Bera (JB):	122.556
Skew:	-0.966	Prob(JB):	2.44e-27
Kurtosis:	5.580	Cond. No.	1.00

Appendix 1.15 – Indicators Outliers

Indicator_value_Deprivation	Indicator_value_Drug_misuse	isOutlier - Indicator_value_Drug_misuse	Indicator_value_Acute_sexually_transmitted_infections	isOutlier - Indicator_value_Acute_sexually_transmitted_infections	Indicator_value_Obese_Children_(Year_6)	isOutlier - Indicator_value_Obese_Children_(Year_6)	Areas
3.9806783036	8.370696171	0	6.9737435378	0	26.91956124	1	Barking and Dagenham LB
3.2450492601	14.82927088	0	7.459728943	1	22.30538922	0	Camden LB
3.7967277136	10.15706806	0	7.0077662627	0	24.92741601	1	Greenwich LB
4.3902719739	12.74835148	0	7.7560658822	1	27.14285714	1	Hackney LB
3.2993892064	11.3423393	0	7.5694232042	1	25.82368655	1	Hammersmith and Fulham LB
4.0593507436	10.27647093	0	7.6016369867	1	23.76367615	0	Haringey LB
3.9845841047	15.38857264	0	7.5369375046	1	22.08121827	0	Islington LB
3.1925420092	13.31345826	0	7.4106121655	1	22.39336493	0	Kensington and Chelsea LB
3.6295509583	11.98071	0	8.0742517633	1	23.97937258	0	Lambeth LB
3.6268762388	10.78953789	0	7.2924724005	0	25.04132231	1	Lewisham LB
4.4389215606	11.5600015	0	7.2061631231	0	25.61604585	1	Newham LB
3.6121329912	12.05924237	0	7.6963878772	1	28.46120874	1	Southwark LB
4.2608660381	16.30729759	1	7.5639646809	1	25.09976057	1	Tower Hamlets LB
2.5262835583	7.409975692	0	7.5168517833	1	19.98964267	0	Wandsworth LB
3.185052852	13.8958245	0	7.5556163663	1	24.80127186	1	City of Westminster LB
4.1057619919	14.03108808	0	6.6899392707	0	25.07682852	1	Knowsley MCD
4.1820399133	16.79313894	1	6.9836363677	0	22.6618705	0	Liverpool MCD
3.6538406886	11.5786018	0	7.3145120585	1	25.04310345	1	Newcastle upon Tyne MCD
4.0989461687	10.35245614	0	6.5810418939	0	25.15633883	1	Sandwell MCD
3.8970976492	18.44221106	1	6.7669383672	0	24.27385892	0	Hartlepool UA
4.007166028	26.34361233	1	6.9154611732	0	19.31216931	0	Middlesbrough UA
3.8939238252	21.21728926	1	7.2603358161	0	18.03399852	0	Blackpool UA
3.9622077687	18.88120964	1	6.6611951581	0	22.53875157	0	Kingston upon Hull UA
2.2291510764	5.343439128	0	5.5779956038	1	17.63817578	0	East Riding of Yorkshire UA
3.3809835797	15.56650246	1	6.872623121	0	19.23380727	0	Derby UA
3.2887932553	16.9181186	1	7.0813552001	0	19.07600596	0	Bristol UA
2.8368940549	15.8629232	1	6.9485233722	0	15.42810985	0	Bournemouth UA
3.1429775795	11.8961039	0	7.5162083881	1	15.5128853	0	Brighton and Hove UA
3.039561282	6.149615543	0	5.8680429855	0	24.66793169	1	Copeland CD
2.6125203436	8.291299251	0	7.3687287778	1	16.359447	0	Exeter CD
3.827800871	16.05603689	1	6.8199159341	0	18.63979849	0	Hastings CD
0	6.064950006	0	6.2680558756	0	10.61678463	1	Winchester CD
0	5.31801937	0	6.3262726064	0	11.34242642	1	East Hertfordshire CD
0	4.056003213	0	6.4463377914	0	10.34482759	1	St Albans CD
3.9170731971	18.66849949	1	6.6969081614	0	22.54143646	0	Burnley CD
3.1547976893	2.807870472	0	5.5965850722	1	19.67632027	0	East Lindsey CD
3.3752729227	24.89292443	1	7.0098745425	0	22.899729	0	Lincoln CD
0	3.006978741	0	5.5782361831	1	22.319202	0	South Holland CD
3.377594902	17.1184724	1	7.3667234895	1	18.90495868	0	Norwich CD
1.622208755	4.800564551	0	5.0954790503	1	12.00787402	0	Craven CD
0	1.829949658	0	5.5858209393	1	17.46835443	0	Hambleton CD
0	2.920050235	0	5.5724495661	1	16.49484536	0	Ryedale CD
3.6960062491	17.8350624	1	6.7467677916	0	19.092827	0	Mansfield CD
0	2.785454231	0	6.1777522324	0	10.88631985	1	Rushcliffe CD
2.7457436088	15.82921507	1	7.3633381661	1	19.18665276	0	Oxford CD
0	3.239523372	0	6.1261232905	0	10.69958848	1	Waverley CD